

Categorical Data Embedding

須上英¹, 陳定立²

¹ 台北大學, ² 中央研究院

Abstract

Categorical data characterize samples by "categories", not by "quantities". Such a characterization makes it difficult for analysts to explore categorical data. In this talk, we will propose a method for embedding categorical data into the Euclidean space, so that the relationships between categories and between samples can be directly exhibited. The method is to find the subspace that can mostly discriminate between categories. The subspace can be considered as the underlying feature space behind the categorical observations. Then the samples are assigned according to the closest positions to their categories in the feature spaces. The effectiveness of this method will be evaluated by simulations and a clustering analysis.

Keyword: categorical data, embedding, discriminant analysis